Open access • Journal Article • DOI:10.3982/ECTA14138

Nonparametric Inference on State Dependence in Unemployment — Source link

Alexander Torgovitsky

Published on: 01 Sep 2019 - Econometrica (John Wiley & Sons, Ltd (10.1111))

Topics: Nonparametric statistics, Unemployment and Survey of Income and Program Participation

SUPPLEMENT TO "NONPARAMETRIC INFERENCE ON STATE DEPENDENCE IN UNEMPLOYMENT"

(*Econometrica*, Vol. 87, No. 5, September 2019, 1475–1505)

ALEXANDER TORGOVITSKY
Department of Economics, University of Chicago

This supplement contains: (i) Discussions of extending the DPO model to discrete outcomes or higher order state dependence; (ii) a brief survey of semiparametric dynamic binary response models; (iii) proofs for the propositions in the main text; (iv) a discussion of the linearity of parameters and assumptions used in the DPO model; (v) a discussion of dimension reduction strategies; (vi) a discussion of estimation and statistical inference; and (vii) additional empirical estimates.

S1. EXTENSION TO DISCRETE OUTCOMES

THE DPO MODEL EXTENDS READILY TO THE CASE where Y_{it} assumes values in $\{0, 1, \ldots, J\}$ with $J > 1$, so that Y_i assumes values in $\mathcal{Y} \equiv \{0, 1, \ldots, J\}^{T+1}$. Applications of such an extension to the dynamics of employment include Magnac (2000) and Prowse (2012), who examined state dependence under finer categorizations (part-time, full-time, etc.) of employment status. Irace (2018) applied the DPO model with $J > 1$ to study the dynamics of hospital choice.

In this more general case, there are $J + 1$ potential outcomes $\{U_{it}(y)\}_{y=0}^{J}$ for each $t \geq 1$. The observed outcome in period t is determined as

$$Y_{it} = \sum_{y=0}^{J} \mathbb{1}[Y_{i(t-1)} = y] U_{it}(y).$$

The primitive P is a probability mass function for $(Y_{i0}, \{U_{it}(0), \ldots, U_{it}(J)\}_{t=1}^{T})$ and the characterization of the identified set remains conceptually unchanged. Some parameters and identifying assumptions that are appropriate for the $J = 1$ case are also appropriate for the $J > 1$ case, but others would require modification. A separate analysis seems beyond the scope of this paper.

S2. EXTENSION TO HIGHER ORDER STATE DEPENDENCE

The discussion in the main text presumes that the analyst is interested in first-order state dependence, that is, the causal effect of the immediately preceding period on the current period. This is consistent with much of the empirical and theoretical literature on state dependence. However, the DPO model can also be modified to consider the causal effects of longer histories of past outcomes. In this section, I outline how one would extend the model to enable the estimation of this type of higher order state dependence.

When Y_{it} is binary, the generalization to state dependence of length $K \geq 2$ is accomplished by introducing 2^K potential outcomes $\{U_{it}(y)\}_{y \in \{0,1\}^K}$ for each period $t \geq K$. The recursive relationship (1) is replaced by

$$Y_{it} = \sum_{y \in \{0,1\}^K} U_{it}(y) \mathbb{1}\big[(Y_{i(t-1)}, \ldots, Y_{i(t-K)}) = y\big] \quad \text{for } t \geq K, \tag{S1}$$

Alexander Torgovitsky: atorgovitsky@gmail.com

 https://doi.org/10.3982/ECTA14138

with the joint determination of periods $t = 0$ up to $t = K - 1$ not being modeled explicitly. For example, with $K = 2$, (S1) would become

$$\begin{aligned} Y_{it} = {} & \mathbb{1}[Y_{i(t-1)} = 0, Y_{i(t-2)} = 0]U_{it}(0,0) + \mathbb{1}[Y_{i(t-1)} = 0, Y_{i(t-2)} = 1]U_{it}(0,1) \\ & + \mathbb{1}[Y_{i(t-1)} = 1, Y_{i(t-2)} = 0]U_{it}(1,0) + \mathbb{1}[Y_{i(t-1)} = 1, Y_{i(t-2)} = 1]U_{it}(1,1), \end{aligned}$$

so that, for each t, there are four potential outcomes corresponding to the four potential two-period histories immediately prior to period t. The primitive P is a probability mass function for the random vector

$$\left(Y_{i0}, Y_{i1}, \ldots, Y_{i(K-1)}, \left\{U_{it}(y) : y \in \{0,1\}^K\right\}_{t=K}^{T}\right).$$

The identified set $\mathcal{P}^\star$ can be characterized through essentially the same argument as for the first-order case.

S3. A BRIEF SURVEY OF SEMIPARAMETRIC IDENTIFICATION IN DBR MODELS

The most common way to implement (8) is to construct a finitely parameterized likelihood function by imposing a parametric distributional assumption (typically normality) for both $(V_{i1}, \ldots, V_{iT})$ and A_i, and by further assuming that these latent variables are independent of X_i, as well as independent of each other. Maximum likelihood estimates of the parameters in (8) can be used with the maintained parametric distributional assumptions to form estimates of causal parameters like those discussed in Section 3. However, consistency of these estimates depends critically on the validity of the parametric assumptions.

Honoré (2002) raised a number of additional criticisms of this reduced form approach. Many of his points apply equally to parametric implementations of structural DC models like (2). One frequently discussed criticism is the treatment of A_i as a normally distributed random effect that is independent of other explanatory state variables. Since (8) is nonlinear, treating A_i as a fixed effect to be estimated leads to the well-known incidental parameters problem when T is small.[1] Nonlinear differencing arguments can be applied in certain cases (Chamberlain (1984, 1985, 2010), Honoré and Kyriazidou (2000), and Bartolucci and Nigro (2010)), but these depend on very specific functional forms and may therefore amplify concerns about misspecification.[2] Honoré and Lewbel (2002) showed that if there exists an exogenous special regressor with a large amount of variation, then A_i can be treated as a fixed effect while also relaxing distributional assumptions in (8). However, their results only identify the parameter coefficients and not causal parameters such as the ATE. Similarly, Pakes and Porter (2016) allowed A_i to be a fixed effect and also removed parametric distributional assumptions on V_{it}, but their partial identification results concern the index parameters, not the causal parameters that constitute the focus of this paper.

Some researchers have developed point and partial identification results for nonparametric models of dynamic binary outcomes that depart from the threshold crossing form of (8) entirely. Instead, these papers maintain assumptions that imply that Y_{it} follows a homogeneous, first-order Markov process, conditional on permanent latent variable, like

[1] Fernández-Val (2009) argued that the bias on the ATE may be relatively small, even for small T, while Carro (2007) argued that the bias can be mitigated by using a modified maximum likelihood estimator.

[2] See also Bonhomme (2012) for related results and a unifying analysis.

A_i, and possibly also on the initial period, Y_{i0}. Under this type of assumption, Kasahara and Shimotsu (2009) and Browning and Carro (2010, 2014) established point identification by imposing the additional assumption that A_i has sufficiently small finite support relative to T, while Hu and Shum (2012) allowed A_i to be continuously distributed, but imposed a high-level completeness condition. These types of conditions on the distribution of unobserved heterogeneity may be difficult to motivate or interpret in applications. In addition, the first-order conditional Markov property assumed by all of these papers may be unattractive in some settings (see Bhuller, Brinch, and Königs (2017) and Section 6 of Browning and Carro (2014)). The DPO model does not maintain this type of first-order conditional Markov property, although it can be imposed if desired; see Section 4.4.

S4. PROOFS

PROOF OF PROPOSITION 1: Observe that if $P \in \mathcal{P}^\star$, then

$$\begin{aligned}
\mathrm{SD}_t^+(P) &= \mathbb{P}_P\big[Y_{i(t-1)} = 0, U_{it}(0) = 0, U_{it}(1) = 1\big] \\
&\quad + \mathbb{P}_P\big[Y_{i(t-1)} = 1, U_{it}(0) = 0, U_{it}(1) = 1\big] \\
&= \mathbb{P}_P\big[Y_{i(t-1)} = 0, Y_{it} = 0, U_{it}(1) = 1\big] + \mathbb{P}_P\big[Y_{i(t-1)} = 1, U_{it}(0) = 0, Y_{it} = 1\big] \\
&= \mathbb{P}[Y_{i(t-1)} = 0, Y_{it} = 0] + \mathbb{P}[Y_{i(t-1)} = 1, Y_{it} = 1] \\
&\quad - \mathbb{P}_P\big[Y_{i(t-1)} = 0, Y_{it} = 0, U_{it}(1) = 0\big] - \mathbb{P}_P\big[Y_{i(t-1)} = 1, U_{it}(0) = 1, Y_{it} = 1\big],
\end{aligned}$$

where the second equality follows because, under (1), $[Y_{i(t-1)} = 0, U_{it}(0) = 0]$ if and only if $[Y_{i(t-1)} = 0, Y_{it} = 0]$, and $[Y_{i(t-1)} = 1, U_{it}(1) = 1]$ if and only if $[Y_{i(t-1)} = 1, Y_{it} = 1]$. The only restrictions implied on the second two terms are

$$\begin{aligned}
0 &\geq -\mathbb{P}_P\big[Y_{i(t-1)} = 0, Y_{it} = 0, U_{it}(1) = 0\big] \geq -\mathbb{P}_P[Y_{i(t-1)} = 0, Y_{it} = 0], \\
0 &\geq -\mathbb{P}_P\big[Y_{i(t-1)} = 1, U_{it}(0) = 1, Y_{it} = 1\big] \geq -\mathbb{P}_P[Y_{i(t-1)} = 1, Y_{it} = 1],
\end{aligned} \tag{S2}$$

and there are no cross-equation restrictions between these terms. Thus, there exists a $P \in \mathcal{P}^\star$ obtaining both of the upper bounds in (S2), and one obtaining both of the lower bounds. The upper and lower bounds in (14) now follow from those in (S2). The bounds in (15) follow from an analogous argument using the decomposition

$$\begin{aligned}
\mathrm{SD}_t^-(P) &= \mathbb{P}[Y_{i(t-1)} = 0, Y_{it} = 1] + \mathbb{P}[Y_{i(t-1)} = 1, Y_{it} = 0] \\
&\quad - \mathbb{P}_P\big[Y_{i(t-1)} = 0, Y_{it} = 1, U_{it}(1) = 1\big] - \mathbb{P}_P\big[Y_{i(t-1)} = 1, U_{it}(0) = 0, Y_{it} = 0\big],
\end{aligned}$$

and the following analogous bounds on the second two terms:

$$\begin{aligned}
0 &\geq -\mathbb{P}_P\big[Y_{i(t-1)} = 0, Y_{it} = 1, U_{it}(1) = 1\big] \geq -\mathbb{P}_P[Y_{i(t-1)} = 0, Y_{it} = 1], \\
0 &\geq -\mathbb{P}_P\big[Y_{i(t-1)} = 1, U_{it}(0) = 0, Y_{it} = 0\big] \geq -\mathbb{P}_P[Y_{i(t-1)} = 1, Y_{it} = 0].
\end{aligned} \tag{S3}$$

Observe that there are no restrictions preventing the terms in (S2) and (S3) from simultaneously achieving either their lower or upper bounds. Considering these two polar cases shows that the sharp identified set for $\mathrm{SD}_t \equiv \mathrm{SD}_t^+ + \mathrm{SD}_t^-$ is equal to $[0, 1]$. *Q.E.D.*

PROOF OF PROPOSITION 2: If $\mathcal{P}^\dagger$ is closed and convex, then $\mathcal{P}^\star$ is also closed and convex, since $\mathcal{P}^\star$ is the set of $P \in \mathcal{P}^\dagger$ that satisfy the linear equalities (11) for all y and x. The image of the continuous, real-valued function θ over this closed, convex, and nonempty set is a closed, nonempty interval (e.g., Rudin (1976, Theorem 4.22)) with smallest value $\theta^\star_{\text{lb}}$ and largest value $\theta^\star_{\text{ub}}$, that is, $\Theta^\star = [\theta^\star_{\text{lb}}, \theta^\star_{\text{ub}}]$. *Q.E.D.*

PROOF OF PROPOSITION 3: Let $P \in \mathcal{P}$ denote a distribution that is consistent with the stated conditions and fix $y \in \{0, 1\}$. Since $\varphi(\bar{s}, \cdot)$ is weakly increasing for each $\bar{s}$, it has a generalized inverse, $\varphi^{-1}(\bar{s}, \cdot)$. By Theorem 3.1 of Fang, Hu, and Joe (1994), $(\tilde{S}_{it}, \tilde{S}_{i(t+t')})$ is decreasing in the concordance ordering as a function of $|t'|$, conditional on $\bar{S}_i$. That is, $\mathbb{P}[\tilde{S}_{it} \geq \tilde{s}_1, \tilde{S}_{i(t+t')} \geq \tilde{s}_2 | \bar{S}_i = \bar{s}]$ is decreasing in $|t'|$ for any $\tilde{s}_1$, $\tilde{s}_2$, and $\bar{s}$. Thus, for any integers t', t'' with $|t'| < |t''|$,

$$\begin{aligned}
&\mathbb{P}_P\big[U_{it}(y) = 1, U_{i(t+t')}(y) = 1\big] \\
&\quad = \mathbb{P}\big[\Delta\mathring{\nu}\big(S_{it}(y)\big) \geq 0, \Delta\mathring{\nu}\big(S_{i(t+t')}(y)\big) \geq 0\big] \\
&\quad = \mathbb{E}\big[\mathbb{P}\big[\varphi(\bar{S}_i, \tilde{S}_{it}) \geq 0, \varphi(\bar{S}_i, \tilde{S}_{i(t+t')}) \geq 0 | \bar{S}_i\big]\big] \\
&\quad = \mathbb{E}\big[\mathbb{P}\big[\tilde{S}_{it} \geq \varphi^{-1}(\bar{S}_i, 0), \tilde{S}_{i(t+t')} \geq \varphi^{-1}(\bar{S}_i, 0) | \bar{S}_i\big]\big] \\
&\quad \geq \mathbb{E}\big[\mathbb{P}\big[\tilde{S}_{it} \geq \varphi^{-1}(\bar{S}_i, 0), \tilde{S}_{i(t+t'')} \geq \varphi^{-1}(\bar{S}_i, 0) | \bar{S}_i\big]\big] \\
&\quad = \mathbb{P}_P\big[U_{it}(y) = 1, U_{i(t+t'')}(y) = 1\big],
\end{aligned} \tag{S4}$$

where the third equality follows because $\varphi(\bar{s}, \cdot)$ is right-continuous (e.g., Embrechts and Hofert (2013, Proposition 1(5))), and the final equality reverses the steps of the first three. As discussed in Appendix S5, (S4) implies Assumption DSC if Assumption ST is satisfied. *Q.E.D.*

PROOF OF PROPOSITION 4: Let $P \in \mathcal{P}$ denote a distribution that is consistent with the stated conditions. If A and B are two events and B occurs with probability strictly between 0 and 1, then $\mathbb{P}[A, B] \geq \mathbb{P}[A]\mathbb{P}[B]$ implies $\mathbb{P}[A|B] \geq \mathbb{P}[A|B^c]$.[3] From this, (19) follows from conditions (i) and (ii) whenever $\mathbb{P}[Y_{i(t-1)} = 1, Y_{i(t-2)} = \tilde{y}] \in (0, 1)$:

$$\begin{aligned}
&\mathbb{P}_P\big[U_{it}(y) = 1 | Y_{i(t-1)} = 1, Y_{i(t-2)} = \tilde{y}\big] \\
&\quad = \mathbb{P}_P\big[\Delta\mathring{\nu}\big(S_{it}(y)\big) \geq 0 | \Delta\mathring{\nu}\big(S_{i(t-1)}(\tilde{y})\big) \geq 0, Y_{i(t-2)} = \tilde{y}\big] \\
&\quad \geq \mathbb{P}_P\big[\Delta\mathring{\nu}\big(S_{it}(y)\big) \geq 0 | \Delta\mathring{\nu}\big(S_{i(t-1)}(\tilde{y})\big) < 0, Y_{i(t-2)} = \tilde{y}\big] \\
&\quad = \mathbb{P}_P\big[U_{it}(y) = 1 | Y_{i(t-1)} = 0, Y_{i(t-2)} = \tilde{y}\big].
\end{aligned}$$ *Q.E.D.*

PROOF OF PROPOSITION 5: Note that Assumption TIV implies that

$$\mathbb{P}[U_{it} = u | Y_{i(0:t-1)}, A_i] = \mathbb{P}[U_{i1} = u | Y_{i0}, A_i] = \mathbb{P}[U_{it'} = u | Y_{i(0:t'-1)}, A_i]$$

almost surely, for any t, t' and $u \in \{0, 1\}^2$. As a consequence, if $t' > t \geq 1$, then

$$\mathbb{P}_P[U_{it'} = u, Y_{i(0:t-1)} = y]$$

[3]This follows because $\mathbb{P}[A \mid B] \equiv \mathbb{P}[A, B]\mathbb{P}[B]^{-1} \geq \mathbb{P}[A]$, while on the other hand, $\mathbb{P}[A \mid B^c] = (\mathbb{P}[A] - \mathbb{P}[A, B])(1 - \mathbb{P}[B])^{-1} \leq \mathbb{P}[A]$.

$$
\begin{aligned}
&= \mathbb{E}\big[\mathbb{P}[U_{it'} = u, Y_{i(0:t-1)} = y | Y_{i(0:t'-1)}, A_i]\big] \\
&= \mathbb{E}\big[\mathbb{1}[Y_{i(0:t-1)} = y]\mathbb{P}[U_{it'} = u | Y_{i(0:t'-1)}, A_i]\big] \\
&= \mathbb{E}\big[\mathbb{1}[Y_{i(0:t-1)} = y]\mathbb{P}[U_{it} = u | Y_{i(0:t-1)}, A_i]\big] \\
&= \mathbb{E}\big[\mathbb{P}[U_{it} = u, Y_{i(0:t-1)} = y | Y_{i(0:t-1)}, A_i]\big] \\
&= \mathbb{P}_P[U_{it} = u, Y_{i(0:t-1)} = y],
\end{aligned}
$$

for any $u \in \{0,1\}^2$ and $y \in \{0,1\}^t$, as claimed. Summing both sides of this equality over all realizations y of $Y_{i(0:t-1)}$ shows that Assumption TIV implies Assumption ST with $m = 0$. *Q.E.D.*

PROOF OF PROPOSITION 6: Let $P \in \mathcal{P}$ denote a distribution that is consistent with the stated conditions. Then, for any $u = (u_0, u_1) \in \{0,1\}^2$,

$$
\begin{aligned}
&\mathbb{P}_P[U_{it} = u | Y_{i(0:t-1)}, \bar{S}_i] \\
&\quad = \mathbb{E}_P\big[\mathbb{P}_P[U_{it} = u | Y_{i(0:t-1)}, \bar{S}_i, \tilde{S}_{i(1:t-1)}] | Y_{i(0:t-1)}, \bar{S}_i\big] \\
&\quad = \mathbb{E}_P\big[\mathbb{P}_P[U_{it} = u | Y_{i0}, \bar{S}_i, \tilde{S}_{i(1:t-1)}] | Y_{i(0:t-1)}, \bar{S}_i\big] \\
&\quad = \mathbb{E}_P\big[\mathbb{P}_P[U_{i1} = u | Y_{i0}, \bar{S}_i] | Y_{i(0:t-1)}, \bar{S}_i\big] = \mathbb{P}_P[U_{i1} = u | Y_{i0}, \bar{S}_i],
\end{aligned}
$$

where the second equality follows because $Y_{i(1:t-1)}$ are fully determined by Y_{i0}, $\tilde{S}_{i(1:t-1)}$, and $\bar{S}_i$ under (7) and (1). The third equality used condition (ii), since when U_{it} is determined by (7), it is a function of $(S_{it}(0), S_{it}(1))$, which have here been split into $(\bar{S}_i, \tilde{S}_{it})$. The time-invariant variable $\bar{S}_i$ serves the role of A_i in the statement of Assumption TIV. *Q.E.D.*

PROOF OF PROPOSITION 7: Let $P \in \mathcal{P}$ denote a distribution that is consistent with the stated conditions and fix a $y \in \{0,1\}$. Consider any (x^0, x^1), $(x^0, \tilde{x}^1)$ in the support of (X_{it}^0, X_{it}^1), with $\tilde{x}^1 \geq x^1$. Then

$$
\begin{aligned}
&\mathbb{P}_P\big[U_{it}(y) = 1 | X_{it}^0 = x^0, X_{it} = x^1\big] \\
&\quad = \mathbb{P}_P\big[\varphi\big(x^0, x^1, V_{it}\big) \geq 0 | X_{it}^0 = x^0, X_{it}^1 = x^1\big] \\
&\quad = \mathbb{P}_P\big[\varphi\big(x^0, x^1, V_{it}\big) \geq 0 | X_{it}^0 = x^0, X_{it}^1 = \tilde{x}^1\big] \\
&\quad \leq \mathbb{P}_P\big[\varphi\big(x^0, \tilde{x}^1, V_{it}\big) \geq 0 | X_{it}^0 = x^0, X_{it}^1 = \tilde{x}^1\big] \\
&\quad = \mathbb{P}_P\big[U_{it}(y) = 1 | X_{it}^0 = x_0, X_{it}^1 = \tilde{x}_1\big],
\end{aligned}
$$

where the second and third equalities used conditions (iii) and (iv). *Q.E.D.*

PROOF OF PROPOSITION 8: Under Assumption MC, the worker's Bellman equation is given by (24). Let $\mathring{\nu}$ denote the worker's per-period objective function, as in (2). Profiling the effort decision for a fixed employment decision y' gives

$$
\begin{aligned}
e^\star\big(S_{it} \parallel y'\big) &\equiv \arg\max_{e' \in \mathcal{E}} \mathring{\nu}\big(y', e', Y_{i(t-1)}, E_{i(t-1)}, V_{it}, A_i\big) \\
&= \arg\max_{e' \in \mathcal{E}} -\kappa\big(y', e', A_i\big) + \delta \mathbb{E}\big[\nu\big(y', e', V_{i(t+1)}, A_i\big) \mid V_{it}, A_i\big], \qquad \text{(S5)}
\end{aligned}
$$

so that the effort decision in period t is a function of the (fixed) current period employment decision, y', the current period wage shock, V_{it}, and time-invariant heterogeneity, A_i. As a consequence, the counterfactual state in period t had the worker chosen y in period $t-1$ is

$$S_{it}(y) \equiv \big(y, e^\star(V_{i(t-1)}, A_i \parallel y), V_{it}, A_i\big),$$

which depends on the hypothesized previous period employment decision, y, the previous and current period wage shocks, $(V_{i(t-1)}, V_{it})$, and time-invariant heterogeneity, A_i.[4] The worker's present-discounted net utility from choosing employment if the previous period's employment choice was y can therefore be written as

$$\Delta\mathring{\nu}\big(S_{it}(y)\big) = \omega\big(y, e^\star(V_{i(t-1)}, A_i \parallel y), A_i, V_{it}\big) - \Delta\kappa(V_{it}, A_i) + \Delta\gamma(V_{it}, A_i), \tag{S6}$$

where $\Delta\kappa$ and $\Delta\gamma$ are shorthand for

$$\begin{aligned}
\Delta\kappa(V_{it}, A_i) &= \big[\kappa\big(1, e^\star(V_{it}, A_i \parallel 1), A_i\big) - \kappa\big(0, e^\star(V_{it}, A_i \parallel 0), A_i\big)\big], \quad \text{and} \\
\Delta\gamma(V_{it}, A_i) &= \delta\mathbb{E}\big[\nu\big(1, e^\star(V_{it}, A_i \parallel 1), V_{i(t+1)}, A_i\big) \\
&\qquad - \nu\big(0, e^\star(V_{it}, A_i \parallel 0), V_{i(t+1)}, A_i\big) | V_{it}, A_i\big].
\end{aligned}$$

Statements (i)–(iv) can now be proven for the generated potential outcomes (7) by using (S6) as follows.

(i) Notice that $(U_{it}(0), U_{it}(1))$ is fully determined by $(\Delta\mathring{\nu}(S_{it}(0)), \Delta\mathring{\nu}(S_{it}(1)))$, and that from (S6), the latter is only stochastic due to $(V_{i(t-1)}, V_{it})$ and A_i. Thus, if Assumption W(a) is satisfied with $m' = 1$, then since A_i is time-invariant, Assumption ST is also satisfied with $m = 0$. More generally, (7) and (S6) imply that $U_{i(t-m:t)}(0)$ and $U_{i(t-m:t)}(1)$ are stochastic only due to $V_{i(t-m-1:t)}$ and A_i, so that Assumption W(a) with any $m' \geq 1$ implies Assumption ST with $m = m' - 1$.

(ii) Under Assumption W(b),

$$\mathbb{E}\big[\nu\big(y', e', V_{i(t+1)}, A_i\big) \mid V_{it}, A_i\big] = \mathbb{E}\big[\nu\big(y', e', V_{i(t+1)}, A_i\big) \mid A_i\big].$$

From (S5), it follows that $e^\star(V_{it}, A_i \parallel y) = e^\star(A_i \parallel y)$ is only stochastic due to A_i. This implies that $S_{it}(y) = (y, e^\star(A_i \parallel y), V_{it}, A_i)$ is only stochastic due to V_{it} and A_i, and that (S6) can be written as

$$\Delta\mathring{\nu}\big(S_{it}(y)\big) = \omega\big(y, e^\star(A_i \parallel y), A_i, V_{it}\big) - \Delta\kappa(A_i) + \Delta\gamma(A_i). \tag{S7}$$

To see that the conditions of Proposition 3 are satisfied, fix a y, take $\bar{S}_i \equiv A_i$, and take $\tilde{S}_{it} \equiv \omega(y, e^\star(A_i \parallel y), A_i, V_{it})$. Then (S7) can be written as

$$\Delta\mathring{\nu}\big(S_{it}(y)\big) \equiv \tilde{S}_{it} - \Delta\kappa(\bar{S}_i) + \Delta\gamma(\bar{S}_i) \equiv \varphi(\bar{S}_i, \tilde{S}_{it}), \tag{S8}$$

which is an increasing and continuous function of $\tilde{S}_{it}$. Since V_{it} and $V_{i(t-1)}$ are independent, conditional on A_i, so too are $\tilde{S}_{it}$ and $\tilde{S}_{i(t-1)}$, conditional on $\bar{S}_i$. Therefore, the stochastic increasing condition of Proposition 3 is also satisfied.

[4] As in Section 2.2, I am assuming here that the solution to this effort decision is unique.

(iii) As shown in (ii), Assumptions MC and W(b) imply that $(U_{it}(0), U_{it}(1))$ is stochastic only due to V_{it} and A_i. Thus, the same argument as in (i) holds with $m' = m$.

(iv) As shown in (ii), Assumptions MC and W(b) imply that the stochastic components of $(S_{it}(0), S_{it}(1))$ are $(V_{it}, A_i) \equiv (V_{it}, \bar{A}_i, Y_{i0})$. Assumption MC further implies that $V_{it}|\{V_{it'}\}_{t'<t}, \bar{A}_i, Y_{i0}$ has the same distribution as $V_{it}|V_{i(t-1)}, \bar{A}_i, Y_{i0}$. By Assumption W(b), the latter has the same distribution as $V_{it}|\bar{A}_i, Y_{i0}$, which by Assumption W(a) has the same distribution as $V_{i1}|\bar{A}_i, Y_{i0}$. Thus, the conditions of Proposition 6 are satisfied with $\bar{S}_i \equiv \bar{A}_i$ and $\tilde{S}_{it} \equiv V_{it}$. *Q.E.D.*

PROOF OF PROPOSITION (9): Under Assumptions MC and W(b), the worker's net utility from employment can be written as (S7); see Proposition 8(ii). For shorthand, define

$$\Omega(y, A_i) \equiv \Delta\kappa(A_i) - \Delta\gamma(A_i) - \bar{\omega}\big(y, e^\star(A_i \parallel y), A_i\big).$$

Then, since $\omega(y, e, a, v) = \bar{\omega}(y, e, a) + v$,

$$\begin{aligned}
&\mathbb{P}\big[\Delta\hat{\nu}\big(S_{it}(y)\big) \geq 0, \Delta\hat{\nu}\big(S_{i(t-1)}(\tilde{y})\big) \geq 0 \mid Y_{i(t-2)} = \tilde{y}\big] \\
&\quad = \mathbb{E}\big[\mathbb{P}\big[V_{it} \geq \Omega(y, A_i), V_{i(t-1)} \geq \Omega(\tilde{y}, A_i) \mid Y_{i(t-2)} = \tilde{y}, A_i, \{V_{is}\}_{s\leq t-2}\big] \mid Y_{i(t-2)} = \tilde{y}\big] \\
&\quad = \mathbb{E}\big[\mathbb{P}\big[V_{it} \geq \Omega(y, A_i), V_{i(t-1)} \geq \Omega(\tilde{y}, A_i) \mid A_i, \{V_{is}\}_{s\leq t-2}\big] \mid Y_{i(t-2)} = \tilde{y}\big] \\
&\quad = \mathbb{E}\big[\mathbb{P}\big[V_{it} \geq \Omega(y, A_i) \mid A_i\big] \\
&\qquad \times \mathbb{P}\big[V_{i(t-1)} \geq \Omega(\tilde{y}, A_i) \mid A_i\big] \mid Y_{i(t-2)} = \tilde{y}\big],
\end{aligned} \tag{S9}$$

where the second equality follows because $Y_{i(t-2)}$ is fully determined by A_i and $\{V_{is}\}_{s\leq t-2}$ when A_i includes the initial conditions, $(Y_{i\underline{T}}, E_{i\underline{T}})$, and the third equality uses Assumption W(b). Since A_i and V_{it} are assumed to be independent,

$$\mathbb{P}\big[V_{it} \geq \Omega(y, A_i) \mid A_i = a\big] = \mathbb{P}\big[V_{it} \geq \Omega(y, a)\big] \equiv F_t\big(\Omega(y, a)\big),$$

so that (S9) can be written as

$$\begin{aligned}
&\mathbb{P}\big[\Delta\hat{\nu}\big(S_{it}(y)\big) \geq 0, \Delta\hat{\nu}\big(S_{i(t-1)}(\tilde{y})\big) \geq 0 \mid Y_{i(t-2)} = \tilde{y}\big] \\
&\quad = \mathbb{E}\big[F_t\big(\Omega(y, A_i)\big)F_{t-1}\big(\Omega(\tilde{y}, A_i)\big) \mid Y_{i(t-2)} = \tilde{y}\big].
\end{aligned}$$

Below, it will be shown that $F_t(\Omega(y, a))$ and $F_{t-1}(\Omega(\tilde{y}, a))$ are both increasing functions of a. This implies that the covariance between $F_t(\Omega(y, A_i))$ and $F_{t-1}(\Omega(\tilde{y}, A_i))$ is positive conditional on $Y_{i(t-2)} = \tilde{y}$ (or on any other event); see, for example, Lehmann (1966). Thus,

$$\begin{aligned}
&\mathbb{P}\big[\Delta\hat{\nu}\big(S_{it}(y)\big) \geq 0, \Delta\hat{\nu}\big(S_{i(t-1)}(\tilde{y})\big) \geq 0 \mid Y_{i(t-2)} = \tilde{y}\big] \\
&\quad \geq \mathbb{E}\big[F_t\big(\Omega(y, A_i)\big) \mid Y_{i(t-2)} = \tilde{y}\big]\mathbb{E}\big[F_{t-1}\big(\Omega(\tilde{y}, A_i)\big) \mid Y_{i(t-2)} = \tilde{y}\big].
\end{aligned} \tag{S10}$$

By reversing the previous arguments, one has

$$\begin{aligned}
\mathbb{E}\big[F_t\big(\Omega(y, A_i)\big) \mid Y_{i(t-2)} = \tilde{y}\big] &= \mathbb{P}\big[\Delta\hat{\nu}\big(S_{it}(y)\big) \geq 0 \mid Y_{i(t-2)} = \tilde{y}\big], \quad \text{and} \\
\mathbb{E}\big[F_{t-1}\big(\Omega(\tilde{y}, A_i)\big) \mid Y_{i(t-2)} = \tilde{y}\big] &= \mathbb{P}\big[\Delta\hat{\nu}\big(S_{i(t-1)}(\tilde{y})\big) \geq 0 \mid Y_{i(t-2)} = \tilde{y}\big],
\end{aligned}$$

which upon substitution into (S10) implies that $\Delta\mathring{\nu}(S_{it}(y))$ and $\Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y}))$ are positively quadrant dependent, locally at $(0,0)$ and conditional on $Y_{i(t-2)} = \tilde{y}$. The result then follows from Proposition 4.

It remains to be shown that $F_t(\Omega(y,a))$ and $F_{t-1}(\Omega(\tilde{y},a))$ are both increasing functions of a. By definition, both $F_t(\cdot)$ and $F_{t-1}(\cdot)$ are decreasing functions, so it suffices to show that $\Omega(y,a)$ is also a decreasing function of a. For this, we remove the search effort decision from the model (condition (iii)), under which

$$\Omega(y,a) = -\Delta\gamma(a) - \bar{\omega}(y,a).$$

The second term is decreasing in a by condition (iv). As for the first term, under the given assumptions, we have

$$\Delta\gamma(a) = \delta\mathbb{E}\big[\nu(1, V_{i(t+1)}, a) - \nu(0, V_{i(t+1)}, a)\big],$$

so to show that it is increasing in a (and therefore that $-\Delta\gamma(y)$ is decreasing in a), it suffices to show that $\nu(y,v,a)$ has the increasing differences (or supermodularity in this simple setting) property in (y,a), that is, that $\nu(1,v,a) - \nu(0,v,a)$ is increasing as a function of a for all v. Under the maintained assumptions, the Bellman equation is

$$\nu(y,v,a) = \max_{y' \in \{0,1\}} \big\{ y'\big(\bar{\omega}(y,a) + v\big) + \delta\mathbb{E}\big[\nu\big(y', V_{i(t+1)}, a\big)\big]\big\},$$

so that $\nu(y,v,a)$ will have increasing differences in (y,a) for all v if $\tilde{\omega}(y',y,a,v) \equiv y'(\bar{\omega}(y,a) + v)$ has increasing differences in (y',y) for all (a,v), in (y',a) for all (y,v), and in (y,a) for all (y',v); see Proposition 2 of Hopenhayn and Prescott (1992). To see that these conditions are satisfied here, observe that by condition (iv),

$$\tilde{\omega}(1,y,a,v) - \tilde{\omega}(0,y,a,v) = \bar{\omega}(y,a) + v$$

is increasing in both y and a, and

$$\tilde{\omega}\big(y',1,a,v\big) - \tilde{\omega}\big(y',0,a,v\big) = y'\big(\bar{\omega}(1,a) - \bar{\omega}(0,a)\big)$$

is increasing in a. *Q.E.D.*

PROOF OF PROPOSITION 10: Under Assumptions MC and W(b), $\Delta\mathring{\nu}(S_{it}(y))$ is given by (S7). Assumption W(d) then implies that

$$\begin{aligned}
&\mathbb{P}\big[\Delta\mathring{\nu}\big(S_{it}(0)\big) \geq 0\big] \\
&\quad = \mathbb{P}\big[W_{it}(0) \geq \Delta\kappa(A_i) - \Delta\gamma(A_i)\big] \\
&\quad = \mathbb{E}\big[\mathbb{P}\big[W_{it}(0) \geq \Delta\kappa(A_i) - \Delta\gamma(A_i) \mid A_i\big]\big] \\
&\quad \leq \mathbb{E}\big[\mathbb{P}\big[W_{it}(1) \geq \Delta\kappa(A_i) - \Delta\gamma(A_i) \mid A_i\big]\big] = \mathbb{P}\big[\Delta\mathring{\nu}\big(S_{it}(1)\big) \geq 0\big],
\end{aligned}$$

which implies Assumption MATR when $U_{it}(y)$ is determined by (7). Assumption MTR follows similarly, since under Assumption W(e),

$$\mathbb{P}\big[\Delta\mathring{\nu}\big(S_{it}(1)\big) - \Delta\mathring{\nu}\big(S_{it}(0)\big) \geq 0\big] = \mathbb{P}\big[W_{it}(1) \geq W_{it}(0)\big] = 1. \qquad \textit{Q.E.D.}$$

S5. LINEARITY OF PARAMETERS AND ASSUMPTIONS

The parameters and assumptions discussed in the main text can be represented as linear functions of $P = \{P(u, x) : u \in \mathcal{U}, x \in \mathcal{X}\}$. This section demonstrates this point. For notational simplicity, I assume throughout that X_i is degenerate, but it is straightforward to adjust the conditions to allow for X_i to be random by simply conditioning and then averaging over all realizations of X_i. (Alternatively, it is also straightforward to modify the parameters so that they are conditional on certain realizations of X_i.)

First, consider SD_t^+, which can be written as

$$\mathrm{SD}_t^+(P) \equiv \mathbb{P}_P\big[U_{it}(0) = 0, U_{it}(1) = 1\big] = \sum_{u \in \mathcal{U}_t^+} P(u),$$

where $\mathcal{U}_t^+$ is the set of $u = (u_0, u(0), u(1)) \in \mathcal{U}$ such that $u_t(0) = 0$ and $u_t(1) = 1$. This is a linear function of P. To see that $\mathrm{SD}_t^+(\cdot|0)$ is linear, write it as

$$\mathrm{SD}_t^+(P|0) = \frac{\mathbb{P}_P\big[U_{it}(0) = 0, U_{it}(1) = 1, Y_{it} = 0\big]}{\mathbb{P}[Y_{it} = 0]} = \frac{\sum_{u \in \mathcal{U}_t^+(0)} P(u)}{\mathbb{P}[Y_{it} = 0]},$$

where $\mathcal{U}_t^+(0)$ is the set of $u \in \mathcal{U}$ such that $u_t(0) = 0$, $u_t(1) = 1$, and $Y_{it} = 0$ when computed through the recursive relationship (1) with $Y_{i0} = u_0$, $U_{it}(0) = u_t(0)$, and $U_{it}(1) = u_t(1)$. Similar equations follow for $\mathrm{SD}_t^+(P|1)$, $\mathrm{SD}_t^+(P|00)$, and $\mathrm{SD}_t^+(P|11)$.

To demonstrate linearity of the assumptions, consider Assumption MTR, which has the simplest form. Assumption MTR can be written as $\mathbb{P}_P[U_{it}(0) = 1, U_{it}(1) = 0] = 0$ for all $t \geq 1$. Letting $\mathcal{U}_t^{\mathrm{MTR}}$ denote the set of all $u \in \mathcal{U}$ such that $u_t(0) = 1$ and $u_t(1) = 0$, Assumption MTR can also be written as

$$\sum_{u \in \mathcal{U}_t^{\mathrm{MTR}}} P(u) = 0, \tag{S11}$$

for all $t \geq 1$. In terms of the ρ function, this equality constraint can be imposed with two inequalities.[5] Assumptions ST and its variations, TIV, MIV, and MATR, can be imposed similarly by summing over the appropriate subsets of $\mathcal{U}$. Assumption MTS can be imposed using a construction similar to that for $\mathrm{SD}_t^+(P|0)$.

Finally, consider Assumption DSC, which has a different structure. In general, Assumption DSC is a nonlinear restriction, but if Assumption ST holds so that the distribution of $U_{it}(d)$ does not depend on t, then $\mathrm{Corr}_P(U_{it}(d), U_{i(t+s)}(d))$ is decreasing in $|s|$ if and only if $\mathrm{Cov}_P(U_{it}(d), U_{i(t+s)}(d))$ is decreasing in $|s|$. Furthermore, under Assumption ST, the latter is true if and only if $\mathbb{E}_P[U_{it}(d)U_{i(t+s)}(d)]$, that is, $\mathbb{P}_P[U_{it}(d) = 1, U_{i(t+s)}(d) = 1]$, is decreasing in $|s|$. It is straightforward to show that $\mathbb{P}_P[U_{it}(d) = 1, U_{i(t+s)}(d) = 1]$ is a linear function of P using a construction like (S11).

S6. DIMENSION REDUCTION

S6.1. *Computational Considerations*

The optimization problem in Proposition 2 can be quite large. For example, if $T = 6$, then the dimension of the variables in the problem, that is, of $P = \{P(u, x) : u \in \mathcal{U}, x \in \mathcal{X}\}$,

[5] In practice, (S11) is always combined with the requirement that $P(u) \geq 0$, and so it simply reduces to $P(u) = 0$ for all $u \in \mathcal{U}_t^{\mathrm{MTR}}$.

is $2^{2T+1} = 2^{13} = 8192$, even without including any covariates. The number of constraints in the problem—even without any identifying assumptions—is at least $2^{T+1} = 128$ for the observational equivalence conditions (11), plus 2×8192 constraints to ensure that P is contained in the unit interval.

These dimensions are large for an unstructured optimization problem. However, if both ρ and θ are linear so that the problems in Proposition 2 are linear programs, then these dimensions are actually fairly modest. A standard desktop computer with sophisticated linear programming solvers such as CPLEX (IBM (2010)) or Gurobi (Gurobi Optimization (2015)) can finish problems of this size in a matter of seconds. Nevertheless, increasing the length of the panel, T, or including rich, time-varying specifications of covariates both increase the number of variables at an exponential rate. This can quickly become computationally infeasible.

In the remainder of this section, I describe three dimension reduction strategies for addressing this curse of dimensionality. The first strategy combines information across multiple models of shorter time horizons. The second strategy applies a similar construction to the covariates. The third strategy imposes a simple semiparametric structure for the covariates. All three strategies involve a natural and familiar trade-off between the amount of information in the data that is utilized, and the difficulty (both computational and statistical) of harnessing that information. They can be implemented separately or combined together.

S6.2. *Combining Shorter Models*

The most immediate way in which the curse of dimensionality affects the DPO model is through the dimension of the potential outcomes sequence, U_i, which increases exponentially with the time period T. This difficulty is common for models that do not impose a conditional Markov restriction on the observed outcomes. For example, Hyslop (1999) considered a parametric DBR model in which the idiosyncratic error follows an AR(1) process. As observed by Heckman (1981) and Chamberlain (1984), this implies that the observed outcomes are not Markov of any order. The resulting likelihood function for the parametric DBR involves a T-dimensional integral, which is also difficult to approximate when T is large.

In the DPO model, this difficulty can be addressed by constructing several shorter, overlapping models. To see how this works, fix a *model length* $\text{ML} \in \{2, \ldots, T\}$. Then construct a DPO model for the observed sequence $(Y_{it_0}, Y_{i(t_0+1)}, \ldots, Y_{i(t_0+\text{ML})})$ at every initial period $t_0 \in \{0, 1, \ldots, T - \text{ML}\}$. Each of these shorter models relates potential outcomes to observed outcomes through (1) for $t \in \{t_0, \ldots, t_0 + \text{ML}\}$. The case discussed throughout the main text corresponds to setting $\text{ML} = T$.

The primitive object is now a *collection* of probability mass functions $P \equiv \{P_{t_0}\}_{t_0=0}^{T-\text{ML}}$, each of which describes the joint distribution of

$$\big(Y_{it_0}, U_{i(t_0+1)}(0), \ldots, U_{i(t_0+\text{ML})}, U_{i(t_0+1)}(1), \ldots, U_{i(t_0+\text{ML})}(1), X_i\big).$$

Each P_{t_0} should satisfy (10), where $\mathcal{U}$ is now $\{0,1\}^{2\text{ML}+1}$, and each P_{t_0} is restricted to lie in a parameter space $\mathcal{P}_{t_0}^{\dagger}$, that can be specified to satisfy the same types of assumptions discussed in Section 4. The identified set contains all collections $P = \{P_{t_0} : P_{t_0} \in \mathcal{P}_{t_0}^{\dagger}\}_{t_0=0}^{T-\text{ML}}$ of shorter models such that (11) is satisfied for each t_0, and all realizations of $(Y_{it_0}, Y_{i(t_0+1)}, \ldots, Y_{i(t_0+\text{ML})}, X_i)$. The target parameter, θ, is a function of the collection of shorter models, P.

Since P_{t_0} and P_{t_0+1} are distributions that encompass some of the same random variables, the identified set for P must satisfy an additional coherency condition. Mogstad, Torgovitsky, and Walters (2019) described such a condition (in a different model) as *logical consistency*. In the DPO model, the logical consistency condition states that when two overlapping models can both assign a probability to an event, this probability must be the same. That is,

$$\mathbb{P}_{P_{t_0}}[Y_{i(t_0+1)} = y_0, U_{i(t_0+2:t_0+\mathrm{ML})} = u] = \mathbb{P}_{P_{t_0+1}}[Y_{i(t_0+1)} = y_0, U_{i(t_0+2:t_0+\mathrm{ML})} = u]$$
$$\text{for all } (y_0, u) \in \{0,1\}^{1+2(\mathrm{ML}-1)}. \tag{S12}$$

The identified set, $\mathcal{P}^\star$, only contains collections P that satisfy (S12) for all t_0.[6] Intuitively, (S12) aggregates information across the shorter overlapping models.

The benefit of this approach is that it reduces the dimension of variables of optimization from 2^{2T+1} to $(T-\mathrm{ML})2^{2\mathrm{ML}+1}$, which no longer increases exponentially with the length of the panel, T. However, it should also be noted that the coherency condition (S12) constitutes an additional $(T-\mathrm{ML})2^{2(\mathrm{ML}-1)}$ constraints that are not present when $\mathrm{ML} = T$. As a consequence, values of ML close to T may not provide any computational gain, and may in fact be costlier. For values of ML significantly smaller than T, however, the large reduction in the number of variables, combined with a modest increase in the number of constraints, can still net out to massive dimension reduction.

The cost of this approach is the loss of information from modeling less of the distribution of Y_i.[7] This manifests itself in two ways. First, there are fewer observational equivalence conditions. This is because a model of $(Y_{it_0}, Y_{i(t_0+1)}, \ldots, Y_{i(t_0+\mathrm{ML})})$ does not provide a probability for an observable sequence of length greater than $\mathrm{ML}+1$. Second, it may not be possible to impose certain identifying assumptions, such as Assumption ST with $m > \mathrm{ML}-1$, for the related reason that a model of length ML does not provide statements about potential outcome sequences longer than $\mathrm{ML}-1$.

S6.3. *Partitioned Covariates*

Recall that $\mathcal{X}$ denotes the support of the covariates X_i. For each $j = 1, \ldots, J$, let $\mathfrak{X}_j \equiv \{\mathcal{X}_{j1}, \ldots, \mathcal{X}_{jK_j}\}$ denote a finite partition of $\mathcal{X}$ into K_j exhaustive and mutually exclusive sets (or bins), $\mathcal{X}_{jk}$, that are specified by the researcher. Let $X_{ij} \equiv \sum_{k=1}^{K_j} k\mathbb{1}[X_i \in \mathcal{X}_{jk}]$ denote the random variable that takes value k if X_i lands in bin $\mathcal{X}_{jk}$.

Let P_j denote a probability mass function with support contained in $\mathcal{U} \times \mathfrak{X}_j$, and let $\mathcal{P}_j$ denote the set of all such functions that sum to unity, as in (10). Every $P \in \mathcal{P}$ defines a collection of $P_j \in \mathcal{P}_j$ for $j = 1, \ldots, J$ through the relationship

$$P_j : \mathcal{U} \times \mathbb{N} \to [0,1] : P_j(u,k) = \sum_{x \in \mathcal{X}_{jk}} P(u,x). \tag{S13}$$

Given a parameter space, $\mathcal{P}^\dagger$, relationship (S13) generates parameter spaces $\mathcal{P}_j^\dagger$ that each P_j is restricted to lie in. Moreover, if $P \in \mathcal{P}^\star$, then P_j must satisfy the following set of

[6] Proposition 2 and the resulting methodology extend immediately, since these constraints are linear in each P_{t_0}.

[7] Formally, the identified set will be an outer identified set, rather than the sharp identified set.

observational equivalence constraints:

$$\mathbb{P}[Y_i = y, X_i \in \mathcal{X}_{jk}] = \sum_{x \in \mathcal{X}_{jk}} \mathbb{P}[Y_i = y, X_i = x]$$
$$= \sum_{x \in \mathcal{X}_{jk}} \sum_{u \in \mathcal{U}_{\text{oeq}}(y)} P(u, x) = \sum_{u \in \mathcal{U}_{\text{oeq}}(y)} P_j(u, k), \quad \text{for all } y, j \text{ and } k. \quad \text{(S14)}$$

As in the previous section, there is an additional logical consistency constraint that can be imposed across the smaller models $\{P_j\}_{j=1}^J$, namely,

$$\sum_{k=1}^{K_j} P_j(u, k) = \sum_{x \in \mathcal{X}} P(u, x) = \sum_{k=1}^{K_{j'}} P_{j'}(u, k) \quad \text{for any } j, j' \text{ and all } u \in \mathcal{U}. \quad \text{(S15)}$$

As long as the collection $\{P_j\}_{j=1}^J$ is sufficient for evaluating the researcher's target parameter, θ, one can work only with these lower-dimensional objects via (S13)–(S15), rather than with P directly.

To see how this partitioning approach addresses the curse of dimensionality, suppose that $X_i = (X_{i1}, \ldots, X_{id_x})$ and that each X_{ij} has K support points $\{x_{jk}\}_{k=1}^K$. The full distribution P consists of $2^{2T+1} \times J^K$ elements. This can be a large number even if J or K are relatively small. However, suppose that the researcher specifies partitions $\mathfrak{X}_j$ taken as $\mathfrak{X}_j = \{\{x_{1k}\}_{k=1}^K, \ldots, \{x_{Jk}\}_{k=1}^K\}$, so that each partition has K elements corresponding to the jth component of X_i. The total dimension of $\{P_j\}_{j=1}^J$ under this partition is $2^{2T+1} \times JK$, which can be dramatically smaller than $2^{2T+1} \times J^K$. The cost of this approach is a loss of information. This occurs for the same reasons as for the strategy in the previous section: Only a subset of the observational equivalence conditions are being met, and identifying content contained in restrictions that would span across covariate partitions cannot be exploited.

S6.4. *A Semiparametric Specification*

One natural response to the curse of dimensionality is to impose semiparametric restrictions. For doing this, it is more convenient to formulate the DPO model as one of the *conditional* distribution of Y_i given X_i, rather than of their joint distribution. This changes most of the discussion in the paper in only obvious ways; the exception is the discussion of statistical inference in Appendix S7, which would require some reworking. The primitive object of the model changes from a joint distribution to a collection of conditional distributions, written (with mild abuse of notation) as $P = \{P(\cdot|x) : x \in \mathcal{X}\}$, each of which has support contained in $\mathcal{U} \equiv \{0, 1\}^{2T+1}$.

With P as a conditional distribution, a natural semiparametric assumption is that

$$\mathcal{P}^\dagger \subseteq \left\{P \in \mathcal{P} : P(u|x) = h(x)'\beta_u \text{ for some } \beta_u \in \mathbb{R}^{d_h}, \text{ all } u \text{ and } x\right\},$$

where h is a known, vector-valued function of length d_h. This assumption says that for each u, $P(u|x)$ is a linear function of a known transformation of x with coefficient vector β_u.[8] Under this assumption, each P is characterized by a set of parameters $\{\beta_u : u \in \mathcal{U}\}$

[8]Note that unlike a linear probability model, here β_u is still required to be such that $P(u|x) \in [0, 1]$.

that has dimension $2^{2T+1} \times d_h$. This dimension does not depend on the number of support points of X_i, and grows linearly with the dimension of the transformation vector, h, thereby overcoming the curse of dimensionality, while preserving the linear programming structure. The cost is the usual threat of potential misspecification.

When taken to the observational equivalence condition (11), the semiparametric model also implies a lower-dimensional representation for the observed data distribution, since

$$\mathbb{P}_P[Y_i = y|X_i = x] = \sum_{u \in \mathcal{U}_{\text{oeq}}(y)} P(u|x) = h(x)' \left(\sum_{u \in \mathcal{U}_{\text{oeq}}(y)} \beta_u \right) \equiv h(x)' \delta_y. \tag{S16}$$

Thus, it justifies estimating $\mathbb{P}[Y_i = y|X_i = x]$ by a linear probability model. In practice, one would want to ensure that the fitted probabilities are in the unit interval. One way to do this is to estimate a constrained least squares regression

$$\hat{\delta}_y \equiv \underset{\delta \in \mathbb{R}^{d_h}}{\arg\min} \sum_{i=1}^{n} \left(\mathbb{1}[Y_i = y] - h(X_i)'\delta \right)^2 \quad \text{s.t. } 0 \leq h(X_i)'\delta \leq 1 \ \forall i = 1, \ldots, n, \tag{S17}$$

as suggested by Domencich and McFadden (1975, p. 105) or Judge, Griffiths, Hill, Lütkepohl, and Lee (1985, p. 759). Alternatively, one can estimate $\mathbb{P}[Y_i = y|X_i = x]$ using a statistical model chosen for fit, and then run the resulting probabilities through the DPO model.

S7. ESTIMATION AND STATISTICAL INFERENCE

In Section 3, I considered the distribution of observables as if it were known without accounting for any sampling error. As a result, the identified set $\Theta^\star$ was also known without error for a given parameter and given set of assumptions. In this section, I adjust the discussion to account for the statistical variation that arises when modeling the data as an i.i.d. sample from some underlying population distribution. First, I describe how to construct a consistent estimator of $\Theta^\star$. Second, I describe how to construct confidence regions that contain (with probability at least $1-\alpha$) the parameter $\theta_0 = \theta(P_0) \in \Theta^\star$ corresponding to the "true" $P_0 \in \mathcal{P}^\star$ that generated the data. Third, I discuss a specification test that can be used to falsify the hypothesis that such a P_0 exists. Finally, I conduct a Monte Carlo simulation to evaluate these procedures.

S7.1. *The Criterion Function*

Approaches based on direct sample analogs of $\theta^\star_{\text{lb}}$ and $\theta^\star_{\text{ub}}$ are unattractive for two important reasons. First, while these estimators are consistent under weak conditions, their asymptotic distributions are highly nonstandard.[9] Second, sample analogs of $\theta^\star_{\text{lb}}$ and $\theta^\star_{\text{ub}}$ might not exist even when the population identified set is nonempty.[10] Both problems can

[9]See Shapiro and Dentcheva (2014, Chapter 5), who derived the asymptotic distributions of these analog estimators. The results of Andrews and Han (2009) imply that naively bootstrapping or subsampling empirical analogs of $\theta^\star_{\text{lb}}$ and $\theta^\star_{\text{ub}}$ will not lead to valid confidence regions.

[10]Freyberger and Horowitz (2015) studied an instrumental variables model for which the identified set can be represented through the solution to two linear programming problems. They proposed a modified bootstrap procedure based on the sample analogs of the solutions to the linear programs, but their procedure assumes that these sample analogs exist.

be addressed by transforming the characterization of the identified set provided in Proposition 2 into a criterion function, and then using a sample analog of this criterion function as the basis for estimation and statistical inference (e.g., Chernozhukov, Hong, and Tamer (2007)).

Some additional notation is required. Let $\mathcal{W} \equiv \text{supp}(Y_i, X_i)$ denote the joint support of the observable data $W_i \equiv (Y_i, X_i)$. For each $w \equiv (w_y, w_x) \in \mathcal{W} \subset \mathbb{R}^{d_W}$, define

$$m_{\text{oeq},w}(W_i, P) \equiv \mathbb{1}[Y_i = w_y, X_i = w_x] - \sum_{u \in \mathcal{U}_{\text{oeq}}(w_y)} P(u, w_x). \tag{S18}$$

Next, divide the restriction function ρ into a deterministic component ρ_d, and a stochastic component ρ_s with dimension d_s. The deterministic component, $\rho_d : \mathcal{P} \to \mathbb{R}^{d_\rho - d_s}$, is a function defined on $\mathcal{P}$ that does not depend on the distribution of W_i. The stochastic component, $\rho_s : \mathcal{P} \to \mathbb{R}^{d_s}$, is a function defined on $\mathcal{P}$ that is assumed to be representable as a moment condition. That is, it is assumed that there exists a function $m_\rho : \mathcal{W} \times \mathcal{P} \to \mathbb{R}^{d_s}$ for which $\rho_s(P) = \mathbb{E}m_\rho(W_i, P)$. This condition is satisfied by all of the identifying assumptions discussed in Section 4.

For example, Assumption ST would be part of ρ_d, since it is not a restriction that depends on the distribution of observables W_i. On the other hand, Assumption MTS would be part of ρ_s, since it depends on the distribution of $(Y_{i(t-1)}, Y_{i(t-2)})$.

Next, define $\mathcal{P}_d^\dagger \equiv \{P \in \mathcal{P} : \rho_d(P) \geq 0\}$ as the set of deterministic constraints on P. These include not only ρ_d, but also the requirement that $P \in \mathcal{P}$, that is, that P is a probability mass function on $\mathcal{U} \times \mathcal{X}$. Then

$$\begin{aligned} \mathcal{P}^\star = \big\{ P \in \mathcal{P}_d^\dagger : \mathbb{E}m_{\text{oeq},w}(W_i, P) = 0 \ \forall w \in \mathcal{W} \\ \text{and } \mathbb{E}m_{\rho,s}(W_i, P) \geq 0 \ \forall s = 1, \ldots, d_s \big\}, \end{aligned} \tag{S19}$$

where $m_{\rho,s}(W_i, P)$ denotes the sth component of $m_\rho(W_i, P)$. Equation (S19) shows that the DPO model can be viewed as a moment inequality model with parameter space $\mathcal{P}_d^\dagger$, moment equalities $\{\mathbb{E}m_{\text{oeq},w}(W_i, P) = 0\}_{w \in \mathcal{W}}$, and moment inequalities $\{\mathbb{E}m_{\rho,s}(W_i, P) \geq 0\}_{s=1}^{d_s}$. Alternatively and equivalently, let $\eta \in \mathbb{R}_+^{d_s}$ denote a vector of nonnegative slackness variables and define the identified set using only moment equalities as

$$\begin{aligned} \mathcal{R}^\star \equiv \big\{ (P, \eta) \in \mathcal{P}_d^\dagger \times \mathbb{R}_+^{d_s} : \mathbb{E}m_{\text{oeq},w}(W_i, P) = 0 \ \forall w \in \mathcal{W} \\ \text{and } \mathbb{E}m_{\rho,s}(W_i, P) - \eta_s = 0 \ \forall s = 1, \ldots, d_s \big\}. \end{aligned} \tag{S20}$$

Then $\mathcal{P}^\star$ is the projection of the first component of $\mathcal{R}^\star$, that is,

$$\mathcal{P}^\star = \big\{ P \in \mathcal{P} : (P, \eta) \in \mathcal{R}^\star \text{ for some } \eta \in \mathbb{R}_+^{d_s} \big\}.$$

Write the moment functions $\{m_{\rho,s}\}_{s=1}^{d_s}$ and $\{m_{\text{oeq},w}\}_{w \in \mathcal{W}}$ together as $\{m_j\}_{j=1}^{d_m}$ where $d_m = d_s + d_W$ and the first d_s components of $\{m_j\}_{j=1}^{d_m}$ correspond to $\{m_{\rho,s}\}_{s=1}^{d_s}$. A convenient choice of population criterion function is

$$Q(P, \eta) \equiv \sum_{j=1}^{d_s} \big| \mathbb{E}m_j(W_i, P) - \eta_j \big| + \sum_{j=d_s+1}^{d_m} \big| \mathbb{E}m_j(W_i, P) \big|. \tag{S21}$$

Notice that $(P, \eta) \in \mathcal{R}^\star$ if and only if $Q(P, \eta) = 0$ and $(P, \eta) \in \mathcal{P}_d^\dagger \times \mathbb{R}_+^{d_s}$. Using an absolute value loss function instead of the more standard quadratic loss function is computationally convenient for the estimation and inference procedures discussed ahead. Other choices of criterion function are possible in principle, but tend to create computational obstacles.[11] Given an i.i.d. sample $\{W_i\}_{i=1}^n$ of size n, a sample analog of Q is constructed by replacing the population expectation with its (scaled) empirical counterpart:

$$\begin{aligned} Q_n(P, \eta) &\equiv \sum_{j=1}^{d_s} \sqrt{n}\left|\bar{m}_{n,j}(P) - \eta_s\right| + \sum_{j=d_s+1}^{d_m} \sqrt{n}\left|\bar{m}_{n,j}(P)\right|, \quad \text{where} \\ \bar{m}_{n,j}(P) &\equiv \frac{1}{n}\sum_{i=1}^{n} m_j(W_i, P) \quad \text{for } j = 1, \ldots, d_m. \end{aligned} \tag{S22}$$

S7.2. *Estimation*

An estimator of $\Theta^\star$ can be constructed by restricting attention to P that come close to minimizing the sample criterion (S22). Let

$$\bar{Q}_n \equiv \min_{(P,\eta)\in\mathcal{P}_d^\dagger\times\mathbb{R}_+^{d_s}} Q_n(P, \eta) \tag{S23}$$

denote the minimum value of Q_n, and let

$$\mathcal{P}_n \equiv \left\{P \in \mathcal{P}_d^\dagger : Q_n(P, \eta) \leq \bar{Q}_n(1 + \tau_n) \text{ for some } \eta \in \mathbb{R}_+^{d_s}\right\}$$

denote the collection of $P \in \mathcal{P}_d^\dagger$ that yield criterion values within $\tau_n\%$ of the optimum. Then define

$$\begin{aligned} \hat{\theta}_{\text{lb}}^\star &\equiv \min \theta(\mathcal{P}_n) = \min_{(P,\eta)\in\mathcal{P}_d^\dagger\times\mathbb{R}_+^{d_s}} \theta(P) \quad \text{s.t. } Q_n(P, \eta) \leq \bar{Q}_n(1 + \tau_n) \quad \text{and} \\ \hat{\theta}_{\text{ub}}^\star &\equiv \max \theta(\mathcal{P}_n) = \max_{(P,\eta)\in\mathcal{P}_d^\dagger\times\mathbb{R}_+^{d_s}} \theta(P) \quad \text{s.t. } Q_n(P, \eta) \leq \bar{Q}_n(1 + \tau_n). \end{aligned}$$

Theorem S.1 of Mogstad, Santos, and Torgovitsky (2018) provides conditions under which $[\hat{\theta}_{\text{lb}}^\star, \hat{\theta}_{\text{ub}}^\star]$ will be a consistent estimator of $\Theta^\star$ in the Hausdorff metric. The result requires $\Theta^\star$ to be nonempty, so that the model is correctly specified. It also requires $\tau_n \to 0$. For the empirical results in Section 5, I used $\tau_n = 0.25$. This value was chosen because confidence regions constructed using the method discussed in the next section require a similar tuning parameter, and $\tau_n = 0.25$ performed well in terms of size for the Monte Carlo simulations discussed in Appendix S7.6.

[11] In a previous draft of this paper, I used a quadratic criterion function. The finite sample performance was slightly better, but the computation was significantly more difficult for reasons I discuss further in Appendix S7.5. Also, criterion functions that incorporate information on the covariance matrix for the moments may have preferable statistical properties; see Andrews and Soares (2010) and Andrews and Barwick (2012). However, Studentizing the moments introduces nonlinearities when the moments are not additively separable in P. This severely complicates computation, because the constraint sets for the optimization problems proposed ahead become potentially non-convex.

S7.3. *Confidence Regions*

As is common in the literature on inference under partial identification, I will construct confidence regions through test inversion.[12] The tests will be of null hypotheses taking the form $H_0 : t \in \Theta^\star$ for conjectured scalar values t. A natural test statistic for this null is a profiled version of Q_n:

$$\bar{Q}_n(t) \equiv \inf_{(P,\eta)\in\mathcal{P}_d^\dagger(t)\times\mathbb{R}_+^{d_s}} Q_n(P,\eta), \tag{S24}$$

where $\mathcal{P}_d^\dagger(t) \equiv \{P \in \mathcal{P}_d^\dagger : \theta(P) = t\}$. Constructing a confidence region for $\Theta^\star$ by inverting these tests means collecting all t for which $\bar{Q}_n(t)$ is not "too large."

To operationalize such a test, one needs to determine how large is "too large" by approximating the distribution of $\bar{Q}_n(t)$ under the null hypothesis. The asymptotic distribution of $Q_n(P,\eta)$ is itself nonstandard due to the lack of point identification; see, for example, Chernozhukov, Hong, and Tamer (2007), Andrews and Soares (2010), Bugni (2010), and Canay (2010). There is an added difficulty here caused by the infimum in the definition of $\bar{Q}_n(t)$, which is introduced by the desire to conduct profile (or "subvector") inference on $\Theta^\star$ rather than $\mathcal{P}^\star$. The two procedures I consider for this problem are subsampling and the shape restriction approach of Chernozhukov, Newey, and Santos (2015).[13]

The subsampling approach approximates the distribution of $\bar{Q}_n(t)$ under the null hypothesis with the distribution of

$$\bar{Q}_b^{\mathrm{SS}}(t) \equiv \inf_{(P,\eta)\in\mathcal{P}_d^\dagger(t)\times\mathbb{R}_+^{d_s}} Q_b^{\mathrm{SS}}(P,\eta),$$

where $Q_b^{\mathrm{SS}}(P,\eta)$ is defined analogously to $Q_n(P,\eta)$, but constructed instead using a subsample $\{W_i^\star\}_{i=1}^b$ of size b that is randomly drawn (without replacement) from $\{W_i\}_{i=1}^n$. This profiled subsampling procedure was first proposed in Romano and Shaikh (2008); see also Chernozhukov, Hong, and Tamer (2007) and Romano and Shaikh (2010). In the following, I refer to the test that rejects $H_0 : t \in \Theta^\star$ when $\bar{Q}_n(t)$ is greater than the $1-\alpha$ quantile of $\bar{Q}_b^{\mathrm{SS}}(t)$ based on B random subsamples as the SS test. A $1-\alpha$ SS confidence region for $\Theta^\star$ is the set of all t for which the SS test does not reject.

The Monte Carlo simulations in the next section suggest that the SS test can be quite conservative in the DPO model. This leads to low power and excessively wide confidence regions. The procedure for testing shape constraints proposed by Chernozhukov, Newey, and Santos (2015, "CNS") provides an alternative that turns out to be less conservative in the DPO model. Their approach is based on a careful approximation of $\bar{Q}_n(t)$ that takes into account the shape of the constraint set $\mathcal{P}_d^\dagger(t) \times \mathbb{R}_+^{d_s}$.

[12] See Canay and Shaikh (2017) for a recent survey of the literature.

[13] In a previous version of this paper, I also applied the method proposed by Bugni, Canay, and Shi (2017). Monte Carlo results reported in that version of the paper suggest that this approach has low power in the DPO model. Another recently proposed procedure for profile inference in partially identified models is Kaido, Molinari, and Stoye (2016). Unfortunately, their approach is not computationally feasible for the dimension of the nuisance parameters (P) considered here.

To describe their procedure, first define the function

$$Q_n^\star(P,\eta,g,h) \equiv \sum_{j=1}^{d_s}\left|\xi_{n,j}^\star(P)+\frac{1}{n}\sum_{i=1}^{n}\nabla m_j(W_i,P)[g]-h_j\right| + \sum_{j=d_s+1}^{d_m}\left|\xi_{n,j}^\star(P)+\frac{1}{n}\sum_{i=1}^{n}\nabla m_j(W_i,P)[g]\right|.$$

Here, (g,h) are parameters that serve as local deviations to (P,η) and, correspondingly, have dimensions 2^{2T+1} and d_s, respectively. The notation $\nabla m_j(W_i,P)[g]$ stands for the directional derivative of m_j with respect to P, evaluated at P, in the direction g, that is, $\frac{\partial}{\partial\kappa}m_j(W_i,P+\kappa g)|_{\kappa=0}$. The function $\xi_{n,j}^\star$ is defined for each $j=1,\ldots,d_m$ as

$$\xi_{n,j}^\star(P) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[m_j\big(W_i^\star,P\big)-\bar m_{n,j}(P)\big],$$

where $\{W_i^\star\}_{i=1}^n$ is a bootstrap sample drawn i.i.d. with replacement from $\{W_i\}_{i=1}^n$. CNS then approximated the distribution of $\bar Q_n(t)$ with that of

$$\begin{aligned}\tilde Q_n(t) \equiv \min_{(P,\eta,g,h)}\ & Q_n^\star(P,\eta,g,h)\\ \text{s.t. } & (P,\eta)\in\widehat{\mathcal{R}}^\star(t) \text{ and}\\ & (P,\eta)+n^{-1/2}(g,h)\in\mathcal{P}_d^\dagger(t)\times\mathbb{R}_+^{d_s}.\end{aligned} \tag{S25}$$

The second constraint here restricts (g,h) to be small deviations of (P,η) that remain inside the parameter space under the null hypothesis.[14] The first constraint uses the definition

$$\widehat{\mathcal{R}}^\star(t) \equiv \big\{(P,\eta)\in\mathcal{P}_d^\dagger(t)\times\mathbb{R}_+^{d_s} : Q_n(P,\eta)\le\bar Q_n(t)(1+\tau_n)\big\},$$

which is the set of (P,η) that approximately solve (S24). In the Monte Carlo simulations in the next section, I find that $\tau_n=0.25$ works well in terms of size, so this is the value that I use in the application in Section 5.

The distribution of $\tilde Q_n(t)$ can be approximated by redrawing $\{W_i^\star\}_{i=1}^n$ a large number (B) of times and computing $\tilde Q_n(t)$ for each draw. I refer to the test that rejects $H_0: t\in\Theta^\star$ when $\bar Q_n(t)$ is greater than the $1-\alpha$ quantile of these B values of $\tilde Q_n(t)$ as the CNS test. A $1-\alpha$ CNS confidence region for $\Theta^\star$ is the set of all t for which the CNS test does not reject.

CNS provided a set of sufficient conditions under which their procedure controls size uniformly (see their Theorem 6.3). Most of the sufficient conditions are satisfied immediately for the specifications used here because $\mathcal{P}$ is finite-dimensional and compact, and both the moment conditions and constraints on $\mathcal{P}$ are linear in P. This includes CNS's Assumptions 2.1–2.2, 3.2–3.3, 4.1–4.2, 5.1–5.4, 6.1–6.5. Their Assumption 3.1 is satisfied

[14] CNS included an additional tuning parameter that regulates the slackness of inequality constraints in $\mathcal{P}_d^\dagger(t)\times\mathbb{R}_+^{d_s}$. This parameter is theoretically necessary; however, it introduces a non-convexity into (S25) that renders the problem computationally intractable. The Monte Carlo simulations in Appendix S7.6 suggest the CNS test performs well in the DPO model even when this parameter is omitted.

when the sample is drawn i.i.d., and their Assumption 3.4 is satisfied because the objective function is unweighted. Assumption 6.6 is a rate condition on the tuning parameter τ_n and the additional tuning parameter used by CNS that I am excluding for computational considerations; see footnote 14. The sufficient conditions also include a high-level anti-concentration condition (their Assumption 6.7), which is mild in a finite-dimensional setting.

S7.4. *Testing for Misspecification*

An attractive feature of a partial identification strategy is that it provides an immediate specification test based on the nonemptyness of the identified set. Specifically, a rejection of the null hypothesis $H_0 : \mathcal{P}^\star \neq \emptyset$ is evidence of the nonexistence of a $P \in \mathcal{P}^\dagger$ that is consistent with the data, and hence evidence that some of the assumptions embodied in $\mathcal{P}^\dagger$ are false, that is, that the model is misspecified. A natural statistic for such a test is the overall minimum criterion value, $\bar{Q}_n$, defined in (S23). The results of CNS show that one can approximate the distribution of $\bar{Q}_n$ by simulating the distribution of a quantity that is analogous to (S25), but which replaces $\mathcal{P}_d^\dagger(t)$ by $\mathcal{P}_d^\dagger$ throughout. A level-α misspecification test rejects the null that the identified set is nonempty when $\bar{Q}_n$ is greater than the $1-\alpha$ quantile of this simulated distribution. Note that such a test always fails to reject when the estimated identified set is nonempty, since in such cases $\bar{Q}_n = 0$.

S7.5. *Computing Critical Values*

In order to implement the SS and CNS tests, it is important to be able to reliably solve the optimization problems that define $\bar{Q}_n(t)$ and $\tilde{Q}_n(t)$. Reliability—in particular, ensuring that local optima are in fact global optima—is especially important here because each problem needs to be solved a large number of times in the process of resampling and inverting hypothesis tests to construct confidence regions. Both problems are convex as long as each $\bar{m}_{n,j}(P)$ is linear in P for every j and $\mathcal{P}_d^\dagger(t)$ is determined by the intersection of linear equalities and inequalities. These conditions are satisfied by all of the assumptions and parameters discussed in Section 2.

Under these conditions, the optimization problem in the definition of $\bar{Q}_n(t)$ (and hence $\bar{Q}_b^{\mathrm{SS}}(t)$) can be reformulated as a linear program, using a standard reformulation argument for the absolute value function. As a result, solving this problem is not significantly harder than solving the linear programs used to directly estimate the bounds of the identified set. The optimization problem in the definition of $\tilde{Q}_n(t)$ can also be shown to be a linear program, again by reformulating the absolute value function. This is the motivation for choosing the absolute loss function in (S21) rather than a quadratic loss function. With quadratic loss, (S25) would be a (convex) quadratically-constrained quadratic program, due to the definition of $\widehat{\mathcal{R}}^\star(t)$. While such programs can still be solved reliably using widely available solvers, they are significantly more costly to solve than the corresponding problem using an absolute loss function.[15]

[15] In previous drafts of this paper, I used a quadratic loss function and solved the quadratically-constrained quadratic programs. This procedure was substantially more computationally demanding.

TABLE SI

FINITE SAMPLE PROPERTIES OF ESTIMATED BOUNDS IN COLUMN (6) OF TABLE II[a]

		$\hat{\theta}^{\star}_{\text{lb}}$			$\hat{\theta}^{\star}_{\text{ub}}$		
	Sample size	1718	3435	6870	1718	3435	6870
$\text{SD}^{+}_{\text{avg}}$	true	0.034	0.034	0.034	0.933	0.933	0.933
	mean	0.047	0.043	0.040	0.931	0.931	0.932
	std	0.009	0.009	0.007	0.006	0.004	0.003
	rmse	0.015	0.013	0.010	0.006	0.005	0.003
	5/95%	0.031	0.029	0.030	0.940	0.938	0.937
	min/max	0.017	0.020	0.020	0.947	0.942	0.941
$\text{SD}^{+}_{\text{avg}}(\cdot\|0)$	true	0.238	0.238	0.238	0.569	0.569	0.569
	mean	0.319	0.300	0.281	0.546	0.550	0.554
	std	0.067	0.070	0.058	0.029	0.026	0.021
	rmse	0.105	0.093	0.072	0.037	0.032	0.026
	5/95%	0.207	0.193	0.193	0.594	0.591	0.589
	min/max	0.120	0.123	0.119	0.622	0.615	0.605
$\text{SD}^{+}_{\text{avg}}(\cdot\|00)$	true	0.414	0.414	0.414	0.980	0.980	0.980
	mean	0.553	0.521	0.487	0.942	0.948	0.956
	std	0.119	0.125	0.104	0.032	0.030	0.026
	rmse	0.183	0.165	0.127	0.050	0.044	0.035
	5/95%	0.348	0.329	0.328	0.988	0.993	0.994
	min/max	0.193	0.214	0.202	1.00	1.00	1.00
$\text{SD}^{+}_{\text{avg}}(\cdot\|1)$	true	0.016	0.016	0.016	0.963	0.963	0.963
	mean	0.026	0.024	0.021	0.959	0.960	0.961
	std	0.007	0.006	0.005	0.005	0.004	0.003
	rmse	0.012	0.010	0.007	0.006	0.005	0.004
	5/95%	0.016	0.014	0.014	0.966	0.966	0.966
	min/max	0.006	0.009	0.011	0.972	0.969	0.968
$\text{SD}^{+}_{\text{avg}}(\cdot\|11)$	true	0.017	0.017	0.017	0.993	0.993	0.993
	mean	0.027	0.024	0.022	0.989	0.990	0.991
	std	0.007	0.006	0.005	0.004	0.004	0.003
	rmse	0.012	0.010	0.007	0.006	0.005	0.003
	5/95%	0.016	0.015	0.014	0.995	0.995	0.995
	min/max	0.006	0.009	0.011	0.998	0.997	0.997
$\mathbb{P}[\Theta^{\star} = \emptyset$ in sample]		0.882	0.624	0.334	–	–	–

[a]The bounds are computed with $T = 6$ under Assumption ST with $m = 4$. The row 5/95% gives the 0.05 quantile across simulations of $\hat{\theta}^{\star}_{\text{lb}}$ and the 0.95 quantile of $\hat{\theta}^{\star}_{\text{ub}}$. Similarly, the row min/max gives the minimum across simulations of $\hat{\theta}^{\star}_{\text{lb}}$ and the maximum across simulations of $\hat{\theta}^{\star}_{\text{ub}}$. The final row shows the proportion of simulations in which the sample identified set is empty. The statistics are based on 500 replications and the tuning parameter is set at $\tau_n = 0.25$.

S7.6. *Monte Carlo Simulations*

In this section, I report the results of a Monte Carlo study that evaluates the procedures discussed in the preceding sections. The data generating process in the study draws Y_i according to the empirical probabilities in the SIPP sample used in Section 5.

Table SI reports the finite sample properties of the estimated bounds, $\hat{\theta}^{\star}_{\text{lb}}$ and $\hat{\theta}^{\star}_{\text{ub}}$, using the same time horizon as in the application ($T = 6$), and maintaining Assumption ST with $m = 4$ as in column (6) of Table II. The statistics are based on 500 simulation draws, and the tuning parameter τ_n is set to 0.25. Comparing results across increasing sample sizes

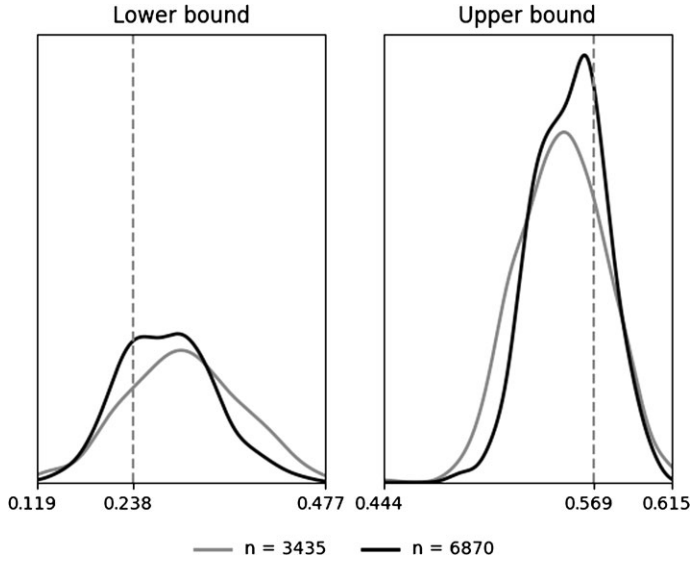

FIGURE S1.—Density of estimated bounds in column (6) of Table II. *Notes*: The plot shows kernel density estimates of the directly estimated lower and upper bounds of $\mathrm{SD}^{+}_{\mathrm{avg}}(\cdot|0)$ from the Monte Carlo reported in Table SI. The gray dotted lines indicate the true lower and upper bounds of the identified set in the DGP. The smoothing used a Gaussian kernel and Silverman's rule-of-thumb bandwidth.

suggests that the bound estimators are consistent. Figure S1 plots the estimated density of $\hat{\theta}^{\star}_{\mathrm{lb}}$ and $\hat{\theta}^{\star}_{\mathrm{ub}}$ when the target parameter is $\mathrm{SD}^{+}_{\mathrm{avg}}(\cdot|0)$. The distributions are nonnormal and nonstandard, which accords with theoretical predictions; see, for example, Section 5 of Shapiro and Dentcheva (2014).

Table SII reports the rejection rates of the SS and CNS tests of $H_0 : t \in \Theta^{\star}$ at nominal levels $\alpha = 0.01$, 0.05, and 0.10 for several values of t. The data generating process is still the SIPP sample, but I only use the first four periods ($T = 3$) in order to moderate the computational burden. The maintained assumptions are Assumption ST with $m = 1$ and Assumption MTR, with the latter assumption also being useful for easing computation. The reported statistics are based on 500 simulation draws, 500 bootstrap draws or subsamples per simulation, and still uses the tuning parameter $\tau_n = 0.25$. The sample size

TABLE SII

FINITE SAMPLE REJECTION PROBABILITIES OF SS AND CNS TESTS[a]

		Rejection probability of $H_0 : t \in \Theta^{\star}$ for $t = \cdots$											
level	test	0.090	0.150	0.210	0.270	0.300	0.338	0.445	0.480	0.510	0.540	0.600	0.660
0.01	SS	0.098	0.032	0.004	0.000	0.000	0.000	0.000	0.002	0.004	0.018	0.140	0.578
	CNS	0.724	0.478	0.242	0.078	0.034	0.012	0.008	0.042	0.146	0.290	0.724	0.982
0.05	SS	0.408	0.190	0.066	0.012	0.006	0.004	0.006	0.014	0.036	0.094	0.394	0.912
	CNS	0.876	0.716	0.472	0.204	0.118	0.048	0.050	0.158	0.326	0.522	0.906	1.00
0.10	SS	0.602	0.394	0.164	0.046	0.016	0.010	0.010	0.032	0.094	0.208	0.588	0.984
	CNS	0.920	0.802	0.596	0.330	0.200	0.092	0.102	0.242	0.430	0.648	0.964	1.00

[a]The target parameter is $\mathrm{SD}^{+}_{\mathrm{avg}}(\cdot|0)$ and Assumptions ST (with $m = 1$) and MTR are maintained. The time horizon is $T = 3$. The values of t in boxes indicate the lower and upper bound of the population identified set. The statistics are based on 500 replications, 500 bootstraps (for CNS) or subsamples (for SS, with $b = n^{2/3} = 228$), and the tuning parameter is set at $\tau_n = 0.25$.

TABLE SIII

FINITE SAMPLE PROPERTIES OF ESTIMATED BOUNDS FROM TABLE SII[a]

		$\hat{\theta}^{\star}_{\text{lb}}$			$\hat{\theta}^{\star}_{\text{ub}}$		
	Sample size	1718	3435	6870	1718	3435	6870
$\text{SD}^{+}_{\text{avg}}$	true	0.037	0.037	0.037	0.925	0.925	0.925
	mean	0.030	0.032	0.034	0.925	0.925	0.925
	std	0.009	0.007	0.005	0.006	0.005	0.003
	rmse	0.011	0.008	0.006	0.006	0.005	0.003
	5/95%	0.014	0.019	0.026	0.934	0.932	0.931
	min/max	0.001	0.010	0.017	0.942	0.938	0.935
$\text{SD}^{+}_{\text{avg}}(\cdot\|0)$	true	0.338	0.338	0.338	0.445	0.445	0.445
	mean	0.274	0.295	0.316	0.457	0.454	0.448
	std	0.075	0.062	0.045	0.040	0.036	0.028
	rmse	0.099	0.075	0.050	0.042	0.037	0.028
	5/95%	0.126	0.172	0.235	0.527	0.520	0.499
	min/max	0.006	0.077	0.168	0.595	0.587	0.548
$\text{SD}^{+}_{\text{avg}}(\cdot\|00)$	true	0.614	0.614	0.614	0.807	0.807	0.807
	mean	0.500	0.537	0.575	0.828	0.821	0.812
	std	0.136	0.115	0.084	0.057	0.051	0.041
	rmse	0.177	0.138	0.093	0.060	0.053	0.041
	5/95%	0.221	0.307	0.422	0.927	0.914	0.882
	min/max	0.011	0.141	0.286	0.962	0.989	0.930
$\text{SD}^{+}_{\text{avg}}(\cdot\|1)$	true	0.016	0.016	0.016	0.957	0.957	0.957
	mean	0.013	0.014	0.015	0.957	0.957	0.957
	std	0.004	0.003	0.003	0.005	0.004	0.003
	rmse	0.005	0.004	0.003	0.005	0.004	0.003
	5/95%	0.006	0.008	0.011	0.964	0.962	0.961
	min/max	0.000	0.004	0.006	0.968	0.967	0.965
$\text{SD}^{+}_{\text{avg}}(\cdot\|11)$	true	0.017	0.017	0.017	0.990	0.990	0.990
	mean	0.014	0.015	0.016	0.990	0.990	0.990
	std	0.004	0.004	0.003	0.003	0.003	0.002
	rmse	0.005	0.004	0.003	0.003	0.003	0.002
	5/95%	0.006	0.008	0.011	0.995	0.994	0.993
	min/max	0.000	0.004	0.007	0.998	0.996	0.996
$\mathbb{P}[\boldsymbol{\Theta}^{\star} = \emptyset$ in sample]		0.428	0.358	0.290	–	–	–

[a]The bounds are computed with $T = 3$ under Assumption ST with $m = 1$ and Assumption MTR. See notes for Table SI.

is set at $n = 3435$, as in the application, and the subsample size is set to $288 \approx n^{2/3}$. The target parameter is taken to be $\text{SD}^{+}_{\text{avg}}(\cdot|0)$. For comparison, Table SIII reports the finite sample performance for the estimated bounds in this case.

The results for t at the boundary of the identified set suggest that the SS test is quite conservative. This leads to low power when testing points outside of the identified set, and thus large confidence regions when constructing these regions through test inversion. In contrast, the CNS test maintains roughly the nominal level at the boundary of the identified set, and is much more powerful at points outside of the identified set. Table SIII suggests that the CNS test may still have low power in this setting; for example, it rejects $t = 0.150$ only about 72% of the time, even though this point is more than three standard deviations smaller than the lower bound of the identified set. These results provide reassurance that the CNS test works reasonably well for the DPO model, at least for the

empirical setting considered here, and provides at least a conservative indication of the impacts of statistical uncertainty.

S8. ADDITIONAL EMPIRICAL RESULTS

Table SIV contains estimates for the SIPP data under some specifications that maintain Assumptions MATR and/or DSC.

TABLE SIV

ADDITIONAL ESTIMATES FROM THE DPO MODEL

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Assumptions				
ST(m)			2	2	2	2	2	2
MTS(q)							2	2
MATR		✓			✓	✓		✓
DSC				✓		✓		✓
				Misspecification				
$\Theta^\star = \emptyset$ in sample	No	No	No	No	No	No	No	No
p-value for $H_0 : \Theta^\star \neq \emptyset$								
				Bounds				
SD^+_{avg}	0.000 0.947	0.000 0.947	0.025 0.933	0.025 0.933	0.027 0.933	0.027 0.933	0.025 0.380	0.027 0.380
$SD^+_{avg}(\cdot\|0)$	0.000 0.581	0.000 0.581	0.193 0.576	0.193 0.576	0.193 0.576	0.193 0.576	0.193 0.560	0.193 0.560
$SD^+_{avg}(\cdot\|00)$	0.000 1.00	0.000 1.00	0.335 0.992	0.335 0.992	0.335 0.992	0.335 0.992	0.335 0.965	0.335 0.965
$SD^+_{avg}(\cdot\|1)$	0.000 0.970	0.000 0.970	0.010 0.966	0.010 0.966	0.010 0.966	0.010 0.966	0.010 0.371	0.010 0.371
$SD^+_{avg}(\cdot\|11)$	0.000 1.00	0.000 1.00	0.010 0.996	0.010 0.996	0.010 0.996	0.010 0.996	0.010 0.383	0.010 0.383
SD_{avg}	0.000 1.00	0.000 1.00	0.054 0.976	0.054 0.976	0.054 0.976	0.054 0.976	0.054 0.423	0.054 0.423

REFERENCES

ANDREWS, D. W. K., AND P. J. BARWICK (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80, 2805–2826. [15]

ANDREWS, D. W. K., AND S. HAN (2009): "Invalidity of the Bootstrap and the m out of n Bootstrap for Confidence Interval Endpoints Defined by Moment Inequalities," *Econometrics Journal*, 12, S172–S199. [13]

ANDREWS, D. W. K., AND G. SOARES (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157. [15,16]

BARTOLUCCI, F., AND V. NIGRO (2010): "A Dynamic Model for Binary Panel Data With Unobserved Heterogeneity Admitting a $\sqrt{n}$-Consistent Conditional Estimator," *Econometrica*, 78, 719–733. [2]

BHULLER, M., C. N. BRINCH, AND S. KÖNIGS (2017): "Time Aggregation and State Dependence in Welfare Receipt," *The Economic Journal*, 127, 1833–1873. [3]

BONHOMME, S. (2012): "Functional Differencing," *Econometrica*, 80, 1337–1385. [2]

BROWNING, M., AND J. M. CARRO (2010): "Heterogeneity in Dynamic Discrete Choice Models," *Econometrics Journal*, 13, 1–39. [3]

——— (2014): "Dynamic Binary Outcome Models With Maximal Heterogeneity," *Journal of Econometrics*, 178, 805–823. [3]

BUGNI, F. A. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78, 735–753. [16]

BUGNI, F. A., I. A. CANAY, AND X. SHI (2017): "Inference for Subvectors and Other Functions of Partially Identified Parameters in Moment Inequality Models," *Quantitative Economics*, 8, 1–38. [16]

CANAY, I. A. (2010): "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity," *Journal of Econometrics*, 156, 408–425. [16]

CANAY, I. A., AND A. M. SHAIKH (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models," in *Advances in Economics and Econometrics*, ed. by B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson. Cambridge University Press, 271–306. [16]

CARRO, J. M. (2007): "Estimating Dynamic Panel Data Discrete Choice Models With Fixed Effects," *Journal of Econometrics*, 140, 503–528. [2]

CHAMBERLAIN, G. (1984): "Chapter 22 Panel Data," in *Handbook of Econometrics*, Vol. 2, ed. by Z. Griliches and M. D. Intriligator. Elsevier, 1247–1318. [2,10]

——— (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer. Cambridge University Press. [2]

——— (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78, 159–168. [2]

CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284. [14,16]

CHERNOZHUKOV, V., W. NEWEY, AND A. SANTOS (2015): "Constrained Conditional Moment Restriction Models," CEMMAP Working Paper CWP 59/15. [16]

DOMENCICH, T., AND D. L. MCFADDEN (1975): *Urban Travel Demand: A Behavioral Analysis*. North-Holland Publishing Co. [13]

EMBRECHTS, P., AND M. HOFERT (2013): "A Note on Generalized Inverses," *Mathematical Methods in Operations Research*, 77, 423–432. [4]

FANG, Z., T. HU, AND H. JOE (1994): "On the Decrease in Dependence With Lag for Stationary Markov Chains," *Probability in the Engineering and Informational Sciences*, 8, 385–401. [4]

FERNÁNDEZ-VAL, I. (2009): "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models," *Journal of Econometrics*, 150, 71–85. [2]

FREYBERGER, J., AND J. L. HOROWITZ (2015): "Identification and Shape Restrictions in Nonparametric Instrumental Variables Estimation," *Journal of Econometrics*, 189, 41–53. [13]

GUROBI OPTIMIZATION (2015): "Gurobi Optimizer Reference Manual." [10]

HECKMAN, J. J. (1981): "Heterogeneity and State Dependence," in *Studies in Labor Markets*, ed. by S. Rosen. University of Chicago Press. [10]

HONORÉ, B. (2002): "Nonlinear Models With Panel Data," *Portuguese Economic Journal*, 1, 163–179. [2]

HONORÉ, B. E., AND E. KYRIAZIDOU (2000): "Panel Data Discrete Choice Models With Lagged Dependent Variables," *Econometrica*, 68, 839–874. [2]

HONORÉ, B. E., AND A. LEWBEL (2002): "Semiparametric Binary Choice Panel Data Models Without Strictly Exogeneous Regressors," *Econometrica*, 70, 2053–2063. [2]

HOPENHAYN, H. A., AND E. C. PRESCOTT (1992): "Stochastic Monotonicity and Stationary Distributions for Dynamic Economies," *Econometrica*, 60, 1387–1406. [8]

HU, Y., AND M. SHUM (2012): "Nonparametric Identification of Dynamic Models With Unobserved State Variables," *Journal of Econometrics*, 171, 32–44. [3]

HYSLOP, D. R. (1999): "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67, 1255–1294. [10]

IBM (2010): "IBM ILOG AMPL Version 12.2," International Business Machines Corporation. [10]

IRACE, M. (2018): "Patient Loyalty in Hospital Choice: Evidence From New York," Working Paper. [1]

JUDGE, G. G., W. E. GRIFFITHS, R. C. HILL, H. LÜTKEPOHL, AND T.-C. LEE (1985): *The Theory and Practice of Econometrics* (Second Ed.). Wiley. [13]

KAIDO, H., F. MOLINARI, AND J. STOYE (2016): "Inference on Projections of Identified Sets," Working Paper. [16]

KASAHARA, H., AND K. SHIMOTSU (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77, 135–175. [3]

LEHMANN, E. L. (1966): "Some Concepts of Dependence," *The Annals of Mathematical Statistics*, 37, 1137–1153. [7]

MAGNAC, T. (2000): "Subsidised Training and Youth Employment: Distinguishing Unobserved Heterogeneity From State Dependence in Labour Market Histories," *The Economic Journal*, 110, 805–837. [1]

MOGSTAD, M., A. SANTOS, AND A. TORGOVITSKY (2018): "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters," *Econometrica*, 86, 1589–1619. [15]

MOGSTAD, M., A. TORGOVITSKY, AND C. R. WALTERS (2019): "Identification of Causal Effects With Multiple Instruments: Problems and Some Solutions," Working Paper. [11]

PAKES, A., AND J. PORTER (2016): "Moment Inequalities for Multinomial Choice With Fixed Effects," Tech. rep. [2]

PROWSE, V. (2012): "Modeling Employment Dynamics With State Dependence and Unobserved Heterogeneity," *Journal of Business & Economic Statistics*, 30, 411–431. [1]

ROMANO, J. P., AND A. M. SHAIKH (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, 138, 2786–2807. [16]

——— (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211. [16]

RUDIN, W. (1976): *Principles of Mathematical Analysis*. New York: McGraw-Hill. [4]

SHAPIRO, A., AND D. DENTCHEVA (2014): *Lectures on Stochastic Programming: Modeling and Theory*, Vol. 16. SIAM. [13,20]

Co-editor Liran Einav handled this manuscript.

Manuscript received 1 February, 2016; final version accepted 30 April, 2019; available online 6 May, 2019.